Path analysis under multicollinearity in S_0 x S_0 maize hybrids

Claudio G. P. de Carvalho[1]; Regiane Borsato[2]; Cosme D. Cruz*[1] and José M. S. Viana[1]

[1] Departamento de Biologia Geral, Universidade Federal de Viçosa, CEP 36571-000 Viçosa, MG, Brazil;

[2] Departamento de Biologia Geral, Universidade Federal do Paraná, CEP 80035-010 Curitiba, PR, Brazil.

(* Corresponding Author. E-mail: cdcruz@mail.ufv.br)

ABSTRACT

The goal of this study was to partition genotypic correlations into direct and indirect yield components effects of grain weight of S_0 x S_0 maize hybrids by means of path analysis. Since multicollinearity was expressed in the independent variable correlation matrix, data adequacy to the analysis was evaluated. Least squares alternative methods were used to avoid the adverse effects of this phenomenon. The experimental design was randomized complete blocks with two replications. The associations among traits obtained from the path analysis under multicollinearity, using least squares solutions, showed little coherence. When using the ridge path analysis or the analysis with trait culling, the traits ear number and 50-kernel weight had the highest direct effects on grain weight/plant. These methods were efficient to reduce the multicollinearity effects.

KEY WORDS: Breeding, ridge regression, yield components, *Zea mays*.

INTRODUCTION

Correlation studies allow for the verification of indirect selection viability in providing genetic gains faster than in direct selection. However, an interpretation of a simple correlation can be the wrong procedure in the selection strategy due to the fact that a high correlation between traits can be a consequence of other traits (Dewey and Lu, 1959).

To understand the causes of trait associations, Wright (1921) proposed the path analysis, which is helpful in partitioning correlation into direct and indirect effects. Ottaviano and Camussi (1981); Ivanovic and Rosic (1984); Singh et al. (1995); Agrama (1996); Djordjevic and Ivanovic (1996) and Arias et al. (1999) studied this methodology in maize (*Zea mays* L.).

It is necessary to obtain path coefficients to estimate direct and indirect effects of a trait group on a basic trait. These coefficients are calculated through regression equations using previously standardized variables. However, coefficient estimates can be adversely affected by multicollinearity between traits. Depending on the degree of multicollinearity, the variances associated with path coefficient estimates can assume high but less reliable values (Carvalho, 1995).

In order to avoid multicollinearity adverse effects, the parameters can be estimated using either trait culling or the least squares alternative method proposed by Carvalho (1995). This method modifies the normal equation system adding a constant K in the diagonal of the independent variable correlation matrix, similar to the ridge regression proposed by Hoerl and Kennard (1970a).

The objectives of this study were: (i) to partition genotypic correlations into direct and indirect yield component effects on grain weight of S_0 x S_0 maize hybrids by means of path analysis; (ii) to evaluate data suitability to this analysis using least squares alternative methods to avoid adverse effects of multicollinearity in the independent variable correlation matrix.

MATERIAL AND METHODS

The experiment was carried out in Viçosa, MG, Brazil, with 130 S_0 x S_0 maize hybrids from Flint x Dent crosses. The Flint and Dent composite were synthesized in the Genetic Institute at ESALQ/USP. They were submitted to two mass selection cycles for prolificacy before beginning the recurrent reciprocal selection program. The Dent composite was obtained from the cross among white and yellow dented maize, mainly of the Tuxpeño race, representative of Mexican, Central and South American germplasms. The Flint composite was obtained from the cross among

hard white and yellow maize, representative of Central American, Colombian and Brazilian germplasm.

The experimental design was randomized complete blocks with two replications. Each plot included one 10.2 meters long row, spaced one meter apart. Two seeds/hole were sowed with 0.3 meters between holes in the rows. After 40-45 days, thinning was done, and one plant/hole remained, totalling 36 plants/row.

The assessed traits were: a) grain weight/plant - GW (kg/10.2 m^2), b) ear number - EN, c) 50-kernels weight - 50KW (g), d) number of leaves above the first ear - LA, e) number of leaves below the first ear - LU, f) plant height - PH (m), g) ear height - EH (m), h) root lodging - RL, i) stalk lodging - SL and j) days to flowering - DF.

Genotypic correlation estimates (r) were obtained according to Mode and Robinson (1959). The variance used to test the significance level of r was calculated as proposed by Vencovsky and Barriga (1992). The correlations were partitioned into direct and indirect effects by means of path analysis (Wright, 1921). In the regression model established, GW was the dependent variable and the other traits were considered independent.

The correlation matrix among independent variables had at least one value close to one. This condition is enough to cause problems in the regression analysis due to multicollinearity (Kmenta, 1971). Path coefficients were estimated discarding the traits that contributed most to the phenomenon. It was also used an alternative method to least squares, proposed by Carvalho (1995). This method adopts the following equation: $(X'X + KI_P)b^* = X'Y$

where $X'X$ is the correlation matrix among independent variables of the regression model; K is a small amount added to $X'X$ matrix diagonal; is the identity matrix (p = parameter number); b^* is the ridge path coefficient vector; and $X'Y$ is the correlation matrix between the dependent variable of the regression model and each independent. The value of the constant K was determined by the ridge trace exam (Hoerl and Kennard, 1970a,b). Ridge trace appeared as a two-dimensional chart, showing how the path coefficient values vary as a function of K in the interval $0< K <1$. K was the smallest value capable of stabilizing most of the path coefficients.

The multicollinearity degree of the $X'X$ matrix was established on the basis of its number of condition (NC = ratio between higher and smallest positive eigenvalues of the matrix) (Montgomery and Peck, 1981). Multicollinearity did not cause serious problems to the analysis when NC<100. Multicollinearity was considered moderate to strong when 100<NC<1000, and severe when NC>1000 . The eigenvalue analysis was also used to identify linear dependency causes between traits to determine those which contributed to multicollinearity. Traits that smallest the highest elements of eigenvectors associated to the smallest eigenvalues, were those which contributed more to multicollinearity (Belsley et al., 1980). Multicollinearity diagnosis, as all the other analysis, was carried out by the GENES software(Cruz, 1997).

RESULTS AND DISCUSSION

Significant differences (P<0.01) were obtained among the S_0 x S_0 maize hybrids for all the traits assessed (Table 1). The coefficients of variation varied from 4.42 to 16.62 for GW, EN, 50KW, PH and EH. These values are considered low or medium according to classification by Scapim et al. (1995), which indicates good experimental accuracy. The variation coefficients varied from 2.12 to 70.42 for the traits LA, LU, RL, SL and DF. Although there is no appropriate classification in the literature for the coefficients of variation associated with these traits in maize, similar values were obtained by Soares (1979) and Ferrão (1985).

Genotypic correlations among traits are shown in Table 2. Ear number (r = 0.67), 50KW (r = 0.30) and LA (r = 0.30) showed highest correlations with GW. Path analysis was performed to verify if correlation magnitudes represented cause and direct effect or indirect effect of other traits (Table 3).

Table 1 - Analysis of variance, mean and coefficient of variation of ten traits assessed in a S_0 x S_0 maize hybrid experiment.

F.V.	G.L.	Q.M.									
		GW[1/]	EN	50KW	LA	LU	PH	EH	RL	SL	DF
Blocks	1	-	-	-	-	-	-	-	-	-	-
Hybrids	129	2.346**	68.651**	4.434**	0.987**	0.406**	0.043**	0.037**	11.465**	19.425**	7.629**
Residue	130	1.209	28.220	0.990	0.275	0.078	0.015	0.014	4.520	6.150	2.817
Mean		6.616	39.56	16.23	8.54	6.82	2.72	1.84	3.02	5.28	79.02
C.V. (%)		16.62	13.43	6.11	6.14	4.10	4.51	6.42	70.42	47.01	2.12

** indicate significance at P<0.01.

[1/] GW - grain weight/plant, EN - ear number, 50KW - 50-kernels weight, LA - number of leaves above the first ear, LU - number of leaves below the first ear, PH - plant height, EH - ear height, RL - root lodging, SL - stalk lodging, DF - days to flowering.

Table 2 - Genotypic correlation estimates among ten traits assessed in maize S_0 x S_0 hybrids.

Traits[1/]	EN	50KW	LA	LU	PH	EH	RL	SL	DF
GW	0.67**	0.30*	0.30*	0.11	0.21*	0.21*	0.21*	0.04	-0.22*
EN		-0.18	0.42**	0.14	0.29*	0.40**	0.08	0.17	-0.15
50KW			0.04	-0.23*	-0.24*	0.01	0.36*	0.01	-0.20*
LA				0.22*	0.53**	0.72**	0.31*	0.19	0.34*
LU					0.15	0.03	0.22*	0.01	-0.03
PH						0.94**	0.77**	0.39**	0.18*
EH							0.61**	0.33*	0.34*
RL								0.44**	0.45**
SL									0.31*
DF									

* and ** indicate significance at P<0.05 and P<0.01, respectively.

[1/] For abbreviations see Table 1.

When using the least squares method (K = 0) and solving analysis with all the traits assessed, a considerable direct effect of RL on GW (0.691) was obtained. This implies that increasing root lodging helped to increase grain weight. The little coherence of this result was considered a multicollinearity adverse effect.

The NC value (52.77) for the $X'X$ matrix shows light multicollinearity among independent variables. However, when properties of the genotypic correlation matrix were tested, one negative eigenvalue was found. By definition, the correlation matrix is non-negative definite, so its eigenvalues should be zero or positive (Searle, 1971). Furthermore, the matrix determinant was negative (D = -0.0139). Therefore, in spite of NC<100, the negative eigenvalue and the negative matrix determinant show the possibility of having problems in the analysis (Hill and Thompson, 1978) carried out by this study probably due to multicollinearity. The high correlation between PH and EH (0.94) points out the association between these characters. Excluding the PH trait from the $X'X$ matrix, its determinant became positive (D = 0.0148) and only non - negative eigenvalues

Table 3 - Estimates of direct effects of yield components on grain weight/plant (GW) obtained from maize S_0 x S_0 hybrids assessment.

Yield components[1]	Direct effects on GW		
	K = 0 [2]	K = 0.25 [2]	Trait culling (PH)[2]
EN	0.662	0.613	0.825
50KW	-0.008	0.504	0.476
LA	0.373	0.018	-0.007
LU	-0.181	0.098	0.106
PH	-0.213	0.457	-
EH	-0.375	-0.302	-0.13
RL	0.691	-0.178	0.042
SL	-0.135	-0.079	-0.102
DF	-0.353	0.080	0.070

[1] For abbreviations see Table 1.
[2] The path coefficients were estimated using the ridge path analysis (K = 0 or K = 0.25) and with trait culling.

were observed. Negative eigenvalues associated with the correlation matrix have been found by Smith et al. (1962); Parodi et al. (1970); Bhatt (1973) and Soares (1987).

The least squares procedure gives unbiased estimators of minimum variance (Searle, 1971). However, since the $X'X$ matrix is adversely affected by multicollinearity ,better results can be obtained using either the ridge path analysis or trait culling. This occurs because the least squares estimator cannot establish reliable b^* ($K = 0$) estimates. The path coefficient vector is an inverse function of that matrix. If there is perfect multicollinearity between some of the independent variables, the matrix will be singular, that is, there shall not be an inverse matrix (Carvalho et al., 1999a). An infinite number of vectors can be established from the generalized $X'X$ inverses, but none of them will have practical meaning. In this context, these authors argue that as the correlation matrix fits best this condition, the path coefficient estimates become less reliable due to of the an increase of the variances associated with these coefficients.

In the ridge path analysis, as K increases, the path coefficient variances are reduced. However, estimates become more biased (Carvalho, 1995). As a consequence, it's not possible to obtain desirable properties of the estimators for such high value. The chosen K value must be one that reduces estimator variance giving only a small bias. So, the mean square error (variance + bias square) will be smaller than the least squares solution (Hoerl and Kennard, 1970b). By examining the ridge trace, the b estimates were similar for the K values close to zero. However, the stability of this vector was only obtained for k=0.25.

Since there is no statistical test to verify if the mean square error of $b^*(K = 0)$ is lower than the one obtained by the least squares method, it's difficult to decide whether ridge path analysis estimates are better. However, the values showed higher coherence when $K = 0.25$ (Table 3). In this case, the direct effect of RL on GW was relatively low (-0.178). Positive direct effects for PH (0.457) and for 50KW (0.504), and direct effects close to zero for LA (0.018) and DF (0.080) were also found. These results were not obtained with $K = 0$.

In addition to obtain the K value, the ridge trace detects multicollinearity and the variables that most affect it (Draper and Smith, 1981). On Figure 1, variation in parameter estimates on the interval $0<K<1$ indicated multicollinearity. The more the coefficients oscillated, the more

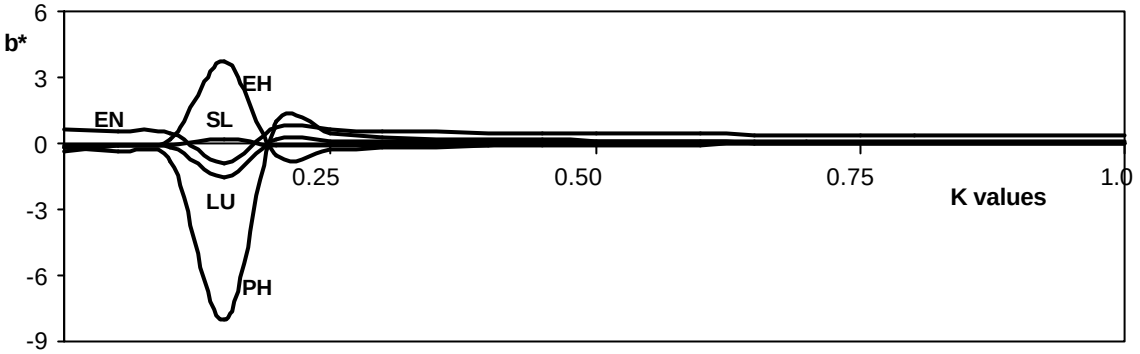

Figure 1 - Path coefficient estimates ($b*$) as a function of K values.

associated variables contributed to the phenomenon. For better demonstration, the figure shows the ridge trace refering to the five traits that contributed most to multicollinearity.

The contribution of each trait to multicollinearity was confirmed by the analysis of the eigenvector elements associated with the lowest eigenvalues. Plant height showed the highest eigenvector element associated with the lowest eigenvalue of the $X'X$ matrix, and ,therefore , contributed more to multi-collinear effects. In general, the discarding/ non-utilization of this trait in the analysis allowed for a path coefficient attainment similar to that of the ridge path analysis (Table 3). The multiple determination coefficient (R^2 = 0.67) for the established regression equation was similar to that of the least squares analysis considering all traits assessed (R^2 = 0.62), and slightly superior to that of the ridge analysis (R^2 = 0.56). Similar results in the path analysis using these alternative methods were obtained with sweet pepper (Carvalho et al., 1999b). In spite of the simplicity of culling traits, Johnston (1972) and Heady and Dillon (1969) describe limitations to the culling of independent variables on regression analyses.

Despite the discrepancy in the results, the three methods indicated positive direct effect (0.613 to 0.825) of EN on GW. In this case, the multivariate analysis on the basis of the path analysis confirmed the result obtained in the simple correlation study. However, this was not verified, e.g., for LA. In table 3, the indirect effects of traits are not observed, and were negligible in general. The low indirect effect of most traits were also found by Singh et al. (1995) in the genetic analysis of maize.

Considering the results described by the ridge path analysis and the analysis with trait culling, it can be seen that EN and 50KW had the highest direct effect on GW. These traits have been shown to influence grain yield in some experiments (Ottaviano and Camussi, 1981; Agrama, 1996; Djordjevic and Ivanovic, 1996; Arias et al., 1999). Although AP showed a direct positive effect on GW, of similar magnitude to that obtained for 50 KW, the feasibility of using this result (increase in AP leading to increase in GW) in the selective process is small. The reason for this is that the maize breeder seeks increase in yield without great increase in the plant height which would imply more lodging and breaking. Furthermore, the relationship of PH with GW depends on EH (indirect effect = - 0.28) and the inclusion of these traits in a multivariate analysis may result in adverse effects caused by multicollinearity in the matrix. The possibility of using this matrix in genetic breeding is discussed by Carvalho (1995) and Carvalho et al. (1999a). Positive and significant direct effect of PH on GW were also estimated on narrow-base maize (Djordjevic and Ivanovic, 1996).

ACKNOWLEDGEMENTS

The authors wish to acknowledge the CAPES for their financial support.

RESUMO

Análise de trilha sob multicolinearidade em híbridos S_0 x S_0 de milho

O objetivo deste trabalho foi desdobrar correlações genotípicas em efeitos diretos e

indiretos de componentes de rendimento sobre peso de grãos de híbridos S_0 x S_0 de milho, por meio da análise de trilha. Adequação dos dados a esta análise foi avaliada, em decorrência da existência de multicolinearidade expressa na matriz de correlações entre os caracteres independentes. Métodos alternativos aos dos mínimos quadrados foram usados para evitar os efeitos adversos deste fenômeno. O delineamento experimental foi o de blocos completos casualizados com duas repetições. As associações entre caracteres obtidas a partir da análise de trilha sob multicolinearidade, utilizando-se as soluções dos mínimos quadrados, foram pouco coerentes. Ao realizar as análises de trilha em crista ou com descarte de variáveis, os caracteres número de espigas e peso de 50 grãos apresentaram os maiores efeitos diretos sobre peso de grãos por planta. Estes métodos foram eficientes para reduzir os efeitos de multicolinearidade.

REFERENCES

Agrama, H.A.S. 1996. Sequential path analysis of grain yield and its components in maize. Plant Breeding. 115: 343–346.

Arias, C.A.A.; Souza Jr., C.L. de and Takeda, C. 1999. Path coefficient analysis of ear weight in different types of progeny in maize. Maydica. 44: 251-262.

Belsley, D.A.; Kuh, E. and Welch, R.E. 1980. Regression diagnostics: identifying data and sources of collinearity. John Wiley and Sons, New York.

Bhatt, G.M. 1973. Significance of path coefficient analysis in determining the nature of character association. Euphytica. 22: 338–343.

Carvalho, S.P. de 1995. Métodos alternativos de estimação de coeficientes de trilha e índices de seleção, sob multicolinearidade. Ph.D. Thesis. UFV, Viçosa.

Carvalho, S.P. de; Cruz, C.D. and Carvalho, C.G.P. de 1999a. Estimating gain by use of a classic selection under multicollinearity in wheat (*Triticum aestivum*). Genetics and Molecular Biology. 22: 109–113.

Carvalho, C.G.P. de.; Oliveira, V.R.; Cruz, C.D. and Casali, V.W.D 1999b. Análise de trilha sob multicolinearidade em pimentão. Pesquisa Agropecuária Brasileira. 34: 603–613.

Cruz, C.D 1997. Programa Genes: Aplicativo computacional em genética e estatística. Editora UFV, Viçosa.

Dewey, D.R. and Lu, K.H. 1959. A correlation and path coefficient analysis of components of crested wheatgrass seed production. Agronomy Journal. 51: 515–518.

Djordjevic, J.S. and Ivanovic, M.R. 1996. Genetic analysis for stalk lodging resistance in narrow-base maize synthetic population ZPS14. Crop Science. 36: 909–913.

Draper, N.R.; Smith, H. 1981. Applied regression analysis. John Wiley and Sons, New York.

Ferrão, M.A.G. 1985. Avaliação do composto de milho (*Zea mays* L.) 'dentado' resultante da seleção recorrente recíproca baseada em famílias de irmãos completos entre os compostos originais 'dentado' e 'duro'. M.S. Diss. Universidade Federal de Viçosa, Viçosa.

Heady, E.O. and Dillon, J.L. 1969. Agricultural production functions. The Iowa State University Press, Ames.

Hill, W.G. and Thompson, R. 1978. Probabilities of non-positive definite between-group or genetic covariance matrices. Biometrics. 34: 429–439.

Hoerl, A.E. and Kennard, R.W. 1970a. Ridge regression: aplications to monorthogonal problems. Technometrics. 12: 69–82.

Hoerl, A.E. and Kennard, R.W. 1970b. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 12: 55–68.

Ivanovic, R.M. and Rosic, K.R. 1984. Path coefficient analysis for three stalk traits and yield in maize. Maydica. 30: 233–239.

Johnston, J. 1972. Econometric methods. McGraw-Hill Book Company, New York.

Kmenta, J. 1971. Elements of econometrics. Macmillan Publishing, New York.

Mode, J.C. and Robinson, H.F. 1959. Pleiotropism and the genetic variance and covariance. Biometrics. 15: 518–537.

Montgomery, D.C. and Peck, E.A. 1981. Introduction to linear regression analysis. John Wiley and Sons, New York.

Ottaviano, E. and Camussi, A. 1981. Phenotypic and genetic relationships between yield components in maize. Euphytica. 30: 601–609.

Parodi, P.; Patterson, F.L. and Nyquist, W.E. 1970. Interrelaciones entre los componentes pricipales y secundarios de rendimiento en trigo (*Triticum aestivum* L.). Fitotecnia Latina Americana. 7: 1–15.

Scapim, C.A.; Carvalho, C.G.P. de and Cruz, C.D. 1995. Uma proposta de classificação dos coeficientes de variação para a cultura de milho. Pesquisa Agropecuária Brasileira. 30: 683-686.

Searle, S.R., 1971. Linear models. John Wiley and Sons, New York.

Singh, G.; Singh, M. and Dhiman, K.R. 1995. Genetic analysis of maize (*Zea mays*) in Sikkim. Indian Journal of Agricultural Sciences. 65: 293–294.

Smith, C.; King, J.W.B. and Gilbert, N. 1962. Genetic parameters of British Large White bacon piggs. Animal Production. 4: 128–143.

Soares, A.A. 1979. Variablidade e correlações genéticas envolvendo o período da polinização à formação da camada preta em grãos de milho (Zea mays L.). M.S. Diss. Universidade Federal de Viçosa, Viçosa.

Soares, P.C. 1987. Correlação, coeficientes de trilha e resposta indireta à seleção em genótipos de arroz (*Oriza sativa* L.) cultivadas em condições de irrigação por inundação contínua e em várzea úmida. M. S. Diss. Universidade Federal de Viçosa, Viçosa.

Vencovsky, R. and Barriga, P. 1992. Genética biométrica no fitomelhoramento. Sociedade Brasileira de Genética, Ribeirão Preto.

Wright, S. 1921. Correlation and causation. Journal of Agricultural Research. 20: 557–585.

Received: July 17, 2000;
Accepted: December 27, 2000.